DOCUMENT RESUME

ED 213 296 HE 014 673

AUTHOR Aubrecht, Judith D.
TITLE Reliability, Validity, and Generalizability of Student Ratings of Instruction. IDEA Paper No. 6.
INSTITUTION Kansas State Univ., Manhattan. Center for Faculty Evaluation and Development in Higher Education.
PUB DATE Nov 81
NOTE 7p.
AVAILABLE FROM Center for Faculty Evaluation and Development, Kansas State University, 1623 Anderson Ave., Manhattan, KS 66502-4098 ($.25 each, minimum order 25 copies).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Faculty; College Instruction; Higher Education; *Instructional Improvement; *Reliability; *Student Evaluation of Teacher Performance; *Teacher Evaluation; Teacher Improvement; *Validity

ABSTRACT

The twin issues of reliability (consistency) and validity (relevance) of student ratings of teacher performance, and the overall issues of generalizability, continue to arise in assessing the usefulness of this faculty evaluation technique. Recent literature addressed four basic questions regarding these issues as they pertain to two faculty evaluation matters, personnel decisions and instructional improvement. The four questions are: (1) Are student ratings biased? (2) Do they agree with the ratings of other groups? (3) Do different kinds of student ratings yield similar information? and (4) Do these ratings help teachers improve instruction? It is neither easy nor very productive to separate the concepts of reliability and validity, and classical reliability theory has evolved into a theory of generalizability, which looks at teacher effect, course effect, student effect, and/or item effect in instruction. Research strengthens the idea that student ratings are both consistent and reliable in teacher evaluation. By and large, strong biasing variables can be controlled statistically, and there is considerable agreement both between student ratings and other groups, and between types of student rating. It also appears that student ratings have served to improve instruction somewhat. More research on generalizability is recommended and appears likely. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

iDEA PAPER No. 6

November, 1981

center for FACULTY EVALUATION & DEVELOPMENT
KANSAS STATE UNIVERSITY

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Kansas State

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED213296

HE 014 673

RELIABILITY, VALIDITY AND GENERALIZABILITY OF STUDENT RATINGS OF INSTRUCTION

Judith D. Aubrecht
Kansas State University

There have been a number of important developments since the publication of IDEA Paper No. 2, "Are Student Ratings of Teacher Effectiveness Valid?" (Aubrecht, 1979). The reader is encouraged to review the previous paper for a discussion of the fundamental issues and definitions. This paper will build on rather than repeat the concepts in IDEA Paper No. 2.

The basic issues surrounding the use of student ratings of instruction have to do with reliability and validity. Reliability studies generally address the question, "Do student ratings of instruction measure consistently whatever it is they measure?" Validity studies address the question, "Do student ratings measure what they are intended to measure for the purposes they serve?" A measure is thus valid to the extent that it is both reliable and relevant. (For example, students could probably report consistently--reliably--on the height of the instructor, but few people would maintain that height is a relevant--and, therefore, valid--measure of teacher effectiveness for use in personnel decisions.)

Validity questions should be addressed within the context of the purposes to be served. There are two primary purposes for administering student ratings; (1) evaluating teachers for personnel decisions and (2) providing information for improvement of instruction.

RELIABILITY

Two types of reliability appear in the literature on student ratings: (1) interrater consistency, and (2) rater stability. A brief discussion of these two concepts appears in footnote 1 of IDEA Paper No. 2 (Aubrecht, 1979).

Interrater Consistency: Suppose the students in a class are considered to be a random sample of all potential students for this class. Then the correlations between the average item responses for this group of students and the average item responses for some other (hypothetical and random) group of students would be the reliabilities of the items. Of course, actual ratings cannot be obtained from "hypothetical" students. However, these correlations can be estimated. One good estimate is based on the average correlation among all pairs of students within the real class. Another estimate comes from dividing the class in half and then correlating the mean responses on each item for the two halves. Since item reliabilities depend on the number of raters, an adjustment is then made to estimate "whole class" reliabilities. The various internal consistency procedures for student ratings show high reliabilities--in the .80's and .90's for classes of twenty or more.

Rater Stability: Stability coefficients are measures of agreement in the ratings of the same students over time. Overall and Marsh (1980) studied student ratings at the end of one hundred courses and again at least a year after the same students had completed their programs. The average stability coefficient was .83. (This study will be cited later in this paper as validity evidence. The reader will see that the issues of reliability and validity are not quite as easily separable as implied above.)

VALIDITY

There are several lines of inquiry in the gathering of evidence to support the validity of student ratings. The basic question is still, "Do student ratings measure what they are intended to measure for the purposes they serve?" There is no single, direct answer to the question. However, answers to complementary subquestions help to illuminate this complex concept. Evidence from each line of inquiry, represented by a question below, is necessary to justify the use of student ratings either for personnel decisions or for instructional improvement.

(1) Are student ratings biased?

(2) Do student ratings agree with the ratings of other groups--teachers, alumni, etc.?

(3) Do different kinds of student ratings yield similar information? Does the information gained from fixed forms, such as IDEA with its predetermined multiple choice items, agree with open-ended responses and/or student interviews?

(4) Do student ratings assist teachers in improving instruction?

In the past two years, research reported on each of these questions has increased our understanding of the fundamental issues and our confidence in the use of student ratings for both personnel decisions and instructional improvement.

ARE STUDENT RATINGS BIASED?

IDEA Paper No. 2 (Aubrecht, 1979) dealt with some of the research in this area, including the impact of initial student motivation and class size. These two variables are not under the control of the teacher (i.e., they are extraneous variables) but they have been shown to affect student ratings. The IDEA system adjusts ("controls") for these effects. Where the biasing effect of variables is systematic, it is possible to eliminate the bias through these statistical adjustments.

The most persistent question of bias has to do with student grades. (See Feldman, 1976.) For example, Brown (1976) found that student grades accounted for about nine percent of the variance in student ratings (on average across classes). An obvious, but not necessarily correct, interpretation of this finding is that teachers can "buy" high student ratings by giving students high grades. Since this is a very popular interpretation of the small but positive relationship between student grades and student ratings typically found in the literature, a further exploration of the controversy is needed.

Howard and Maxwell (1980) suggested a very useful way to illuminate the student grades-student ratings controversy. Figures 1, 2, and 3 below are from their paper. In Figure 1, student grades directly (causally) influence student ratings. That is, students give higher ratings because they are given higher grades.

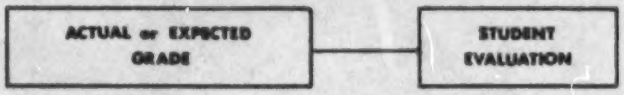

Figure 1. Grading leniency bias model.

Figure 2 suggests an alternative model in which teaching effectiveness influences student performance which in turn influences both student grades and student ratings of instruction. That is, more effective teaching causes more student learning which causes higher grades and higher student ratings.

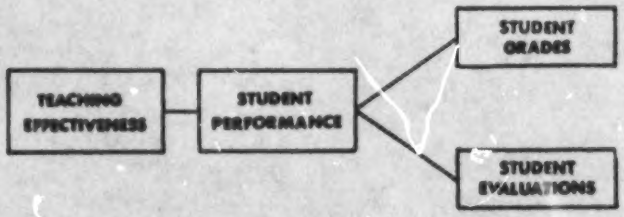

Figure 2. Teaching effectiveness model.

Figure 3 represents another alternative model in which student motivation influences student performance which again influences both student grades and student ratings. That is, higher student motivation causes more student learning which causes higher grades and higher student ratings of instruction. In Figures 2 and 3, there exists a positive relationship between student grades and student ratings. However, this relationship is seen as a legitimate outcome of student performance rather than as a "bias."

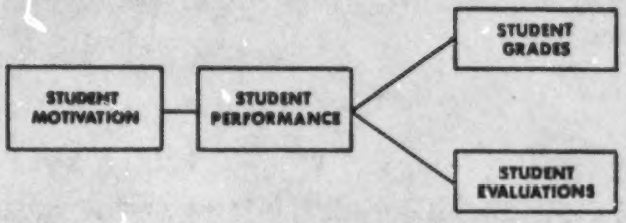

Figure 3. Student characteristics model.

In the same paper, Howard and Maxwell (1980) also reported on two path analytic studies which strongly support the student characteristics (Figure 3 above) model. Path analysis uses correlational evidence to test the appropriateness of causal models. They suggested, but did not test, a model that would combine Figures 2 and 3. This model would assume that both teaching effectiveness and student motivation influence student performance (an idea certainly in accord with most notions of teaching and learning). Student performance in turn would influence student ratings and student grades as in Figures 2 and 3. This combined model also would support a legitimate (not biasing) interpretation of the grades-ratings relationship.

In another path analytic study, Stumpf, Freedman, and Aguanno (1979) examined six possible biasing variables (instructor sex, years of experience, academic rank, proportion of required courses taught, class size, and a publications index.) They found a total of nine percent of the variance in student ratings was accounted for by all of these variables combined. (Note that the strongest of these variables--class size--is controlled in the IDEA system.)

Marsh (1980) used multiple regression analysis to study the effect of sixteen background characteristics--prior subject interest, workload/difficulty, overall GPA, class size, teacher rank, percent of student majors in class, expected grade, reason for taking course (including major requirement, major elective, general interest only, general education requirement, minor/related field), class level (including mean year in school, percentage of freshman and sophomores, percentage of juniors and seniors, upper or lower division). Multiple regression is a way to study the amount of variance in a measure (e.g., student ratings) uniquely accounted for by each of several variables (e.g., the characteristics above). Only four of these characteristics--prior subject interest, workload/difficulty, expected grade, and reason for taking the course (proportion enrolling for general interest)--uniquely accounted for as much as five percent of the variance in any of the evaluative scores. These four variables were then tested through path analysis. Marsh found that:

> As Feldman (1976) anticipated, a sizable portion of the relationship between Expected Grades and student ratings is spurious and attributable to Prior Subject Interest. . . . Averaged across all 10 evaluation scores, the Total Causal Variance estimate was highest for Prior Subject Interest (5.1 percent), followed by Workload/Difficulty (3.6 percent), Expected Course Grade (2.6 percent) and Reason for Taking Course (1.5 percent). (Marsh, 1980, p. 232)

(Note that workload/difficulty is positively related to student ratings. That is, more difficult classes are rated more favorably.)

DO STUDENT RATINGS AGREE WITH THE RATINGS OF OTHER GROUPS?

Student/Instructor: The Marsh, Overall, and Kesler (1979) study cited in IDEA Paper No. 2 was a very important contribution in this area. It demonstrated good absolute agreement and good relative (correlational) agreement between instructors and students on the items in the student rating form. It also used factor analysis to demonstrate that the underlying dimensions of the rating form were basically the same for students and for faculty. Marsh et al. interpreted these findings to mean that students and faculty think of teaching and learning in similar ways.

Student/Teaching Assistant as Instructor: In a paper conceptually related to the Marsh, Overall, and Kesler (1979) study above, Braskamp, Caulley and Costin (1979) studied agreement between ratings by students and self-ratings by relatively inexperienced instructors (seventeen teaching assistants in an introductory psychology class). Student ratings and instructor self-ratings were collected in both fall and spring terms. Again, there was good agreement on items and on underlying factors. However, there were interesting changes from fall to spring. Absolute agreement on items was better in the fall, the first time the ratings were done, than in the spring. (Instructor self-rating decreased in the spring, while student ratings increased.) The extent of agreement on each of the five factors (convergent validity) increased from fall to spring as did the distinctiveness of each of the factors (discriminant validity). The results imply some adjustment by instructors in self-image as well as some conceptual sharpening of ideas about teaching and learning. A comparable study of experienced teachers would be welcome.

Student/Teaching Assistant as Observer: Lamberth and Kosteski (1981) also studied teaching assistants. In this case, the teaching assistants who taught recitation sections attended all the lectures in a large multi-section course. Both teaching assistants and students rated the lecturers. The average item correlations were .71 for one semester and .69 for the next.

Students/Alumni: In the Overall and Marsh (1980) study cited earlier in the section on reliability, the absolute agreement between students and alumni (the same people at least a year after program completion) was very high (median rating of 6.63 for alumni and 6.65 for end-of-term). Wise, Hengstler, and Braskamp (1981) showed that there are almost identical factor structures (underlying dimensions) in student and alumni ratings.

DO DIFFERENT KINDS OF STUDENT RATINGS YIELD SIMILAR INFORMATION?

Ory, Braskamp and Pieper (1980) studied the agreement in student ratings collected in three different ways: objective questionnaire items, written student comments to open-ended questions, and group interview. The average correlation among instructor ratings was .86 and the average correlation among course ratings was .89. The same authors (Braskamp, Ory, and Pieper, 1981) compared written comments to fixed alternative items and found "a high degree of convergent validity." They concluded:

> An average class rating on a fixed alternative item of overall quality of instructor is more cost efficient than a content analysis of student written comments, if overall instructor competence, as judged by students, is a desired piece of evaluative information. (Braskamp, et al., 1981, p. 210)

However, they also recommend collecting written comments if improvement of instruction is one purpose of the evaluation.

DO STUDENT RATINGS ASSIST TEACHERS IN IMPROVING INSTRUCTION?

In IDEA Paper No. 2 (Aubrecht, 1979, p. 4) the answer was, "In all, student ratings taken alone, without assistance in interpretation, seem to be of limited help to a teacher wishing to improve instruction." However, positive results had already been reported with the use of a combination of student ratings and an instructional consultant. More recently, Overall and Marsh (1979) also combined midterm student ratings with instructor/consultant discussions about the use of the ratings in improving instruction. Instructors who received this combined treatment had more favorable end-of-term ratings than others who had not received this midterm feedback. Treatment group instructors had students who performed better on a common final and reported more positive feelings about the subject matter.

The most comprehensive answer to the question comes from Cohen (1980) in a meta-analysis of twenty-two studies on improving teaching through student ratings. (A meta-analysis is a review of the literature that employs statistical techniques to determine a common metric--"an appropriate yardstick"--across relevant studies.) Cohen found an improvement effect for student ratings alone and a larger effect for "augmented student ratings" (i.e., ratings with consultant assistance). In the author's words:

> Thus a typical instructor receiving augmented feedback was performing (at the end of the semester) at the 74th percentile. This can be compared with the typical instructor receiving only student ratings (58th percentile) and the typical instructor receiving no mid-semester feedback (50th percentile). (Cohen, 1980, p. 336)

GENERALIZABILITY

So far this has been a review of the reliability and validity issues related to the analysis of an instructor-course combination ("class"), assessed by a fixed group of students using a given set of items. However, a faculty member will teach a number of different courses to many groups of students over the years. Personnel decisions imply predictions about the quality of instruction in future courses. Hence the term, generalizability, a concept which embodies both reliability and validity.

(1) To what extent do student ratings of past courses taught by an instructor help predict quality of instruction in future courses?

(2) To what extent can we predict the future effectiveness of a course from past student evaluation?

The above questions could be characterized as "predictive validity" questions. They must be accompanied by a series of reliability questions such as:

(1) How many courses taught by a single instructor should be sampled in order to make acceptably accurate predictions about the instructor's future teaching effectiveness?

(2) How many instructors teaching a single course should be sampled to make acceptably accurate predictions about course effectiveness?

Concern about the generalizability of student ratings has been relatively recent among the researchers. Perhaps this vital area could not have been addressed until the more fundamental questions of reliability and validity had been answered with some confidence. Two basic approaches have been taken by the researchers: classical reliability theory and generalizability theory.

Classical reliability theory is used to examine correlations in student ratings under the following conditions: (1) same course-same instructor (different classes of students); hereafter referred to as SC-SI, (2) same course-different instructor; SC-DI, (3) different course-same instructor; DC-SI. The SC-DI condition serves to isolate the "course effect," while the DC-SI condition isolates the "instructor effect." The SC-SI condition gives a "class effect" which is a combination of instructor effect, course effect and an "interaction" between course and instructor. Generalizability theory uses analysis of variance to determine the proportion of variance in student ratings due to "teacher effect," "course effect," and/or "student effect," and/or "item effect."

In the contrast between the two approaches described above, the reader can see that it is no longer easy (nor is it terribly productive) to separate concepts of validity and reliability. This same blurring occurred earlier in this paper in the treatment of the Overall and Marsh (1980) study of student-alumni agreement. The findings of that study were cited as evidence both for stability and for (predictive) validity.

Hogan (1973) used reliability theory to study thirty SC-SI pairs, forty-five DC-SI pairs, and thirty-nine SC-DI pairs. The student rating form used had seven factors. The median correlation for the seven factors was about .70 for the SC-SI pairs and about .40 for the DC-SI pairs. The SC-DI pairs were very inconsistent across the factors with an average correlation of only .19. Hogan suggested from his analysis that about six sets of ratings of a single instructor in different courses are needed to get a .90 reliability for the global rating factor in the instrument used in this study.

In another classical reliability study, Bausell, Schwartz and Purohit (1975) examined similar correlations for items rather than factors. Somewhat lower correlations would be expected for single items rather than factors which are combinations of items. Their results: (1) SC-SI, average item correlation of .64; (2) DC-SI, average item correlation of .37; (3) SC-DI, average item correlation of .24.

Table 1 displays a comparison between the results of the Hogan (1973) study and the Bausell et al. (1975) study. In each cell the first number displayed is the correlation and the second number in parentheses is the percent of the variance accounted for by each of the teacher/course combinations. That is, from Hogan's results 49 percent of the variance in student ratings comes from the instructor-course combination, the "class" effect. Sixteen percent of the variance is "instructor effect" alone, and only four percent is "course effect" alone. (Note that 51 percent of the variance is not accounted for by either instructor or course or the interaction of the two. Student motivation is an example of another variable which would contribute to the variance in student ratings. Also note that the figures for the Bausell et al. (1975) are in general agreement with Hogan's work.

	TEACHER EFFECT	COURSE EFFECT	CLASS EFFECT
HOGAN (1973) STUDY	.40 (16%)	.19 (4%)	.70 (49%)
BAUSELL et al. (1975) STUDY	.37 (14%)	.24 (6%)	.64 (41%)

Table 1
Correlations and Variances in Classical Reliability Studies

Marsh and Overall (1981) used generalizability theory to estimate the amount of variance in student ratings due to teacher effect at 15.5 percent. (Compare Hogan's 16 percent and Bausell's 14 percent.) The course type and level accounted for about five percent of the variance in the same study. (Hogan; four percent and Bausell; six percent.)

Gillmore, Kane, and Naccarato (1978) also used generalizability theory to estimate the amount of variance in student ratings due to teacher effect and to course effect. They conducted two studies. The first isolated a teacher effect from a DC-SI condition. The second isolated a course effect from a SC-DI condition. By combining the two studies they were able to estimate a "teacher-course interaction effect."

It is difficult to compare the Gillmore et al. (1978) work with the earlier reliability studies because Gillmore et al. included in their model additional student and item variables. Table 2 helps to simplify this comparison by looking only at that part of the variance attributable to the "class effect" (the sum of instructor effect, course effect, and instructor-course interaction). A simple calculation on the Hogan (1973) and Bausell et al. (1975) results yields the appropriate data. Hogan reported that 49 percent of the variance in student ratings is "class effect." Of that class effect, 33 percent was teacher effect (16 divided by 49), seven percent was course effect (4 divided by 49), and the rest (60 percent) would have to be teacher-course interaction. A comparable calculation for the Bausell et al. (1975) study yields 34 percent for instructor effect, 15 percent for course effect, and 51 percent for teacher-course interaction. The Gillmore et al. (1978) estimates of variance from the generalizability analysis showed 40 percent of the variance of the "class effect" attributable to teacher effect, six percent to course effect, and 54 percent attributable to teacher-course interaction.

Table 2
Comparison of Results from Classical Reliability and Generalizability Studies: Class Effect

	TEACHER EFFECT	COURSE EFFECT	TEACHER COURSE INTERACTION
HOGAN (1973) STUDY	33%	7%	60%
BAUSELL et al. (1975) STUDY	34%	15%	51%
GILLMORE et al. (1978)	40%	6%	54%

Gillmore et al. (1978) estimated that a sample of five to ten average-sized courses would be needed to yield a dependable measure of teaching effectiveness. (Note the agreement with Hogan's suggested six courses.) They also say, "The results show that if one wishes to evaluate courses by generalizing over teachers, no dependable result is possible for any reasonable sample size of teachers." (p. 9) That is, there is not a sufficiently large and/or stable "course effect" in student ratings to base decisions about courses on such instruments.

One safe prediction: Many more generalizability studies--using either reliability or generalizability theory--will appear during the next few years in the literature on student ratings.

SUMMARY

In IDEA Paper No. 2, this author summarized by suggesting that student ratings are adequately reliable and relevant measures to be included in teacher evaluation procedures. The research since that time serves to strengthen that conclusion. While studies of "bias" are far from over, it is unlikely that anyone will find a strong biasing variable that cannot be controlled statistically. Studies comparing student ratings to the ratings of other groups--alumni, teaching assistants, and the instructors themselves--show considerable correspondence. And there is great agreement when different types of student ratings are compared. Evidence on all of these "research fronts" supports the validity of student ratings for personnel decisions.

Perhaps the best news has been in the research on the use of student ratings to improve instruction. It now seems that there is a small improvement effect related to the use of student ratings alone. However, this effect is greatly enhanced by giving assistance to the teacher both in interpreting the ratings and in other instructional development activities.

The most important questions about student ratings are now in the area of generalizability. Does it really take the rating of five or six different courses taught by a single instructor to predict teaching effectiveness? Do some instruments, or items within instruments, predict better than others? What are the implications for personnel decisions?

REFERENCES

Aubrecht, J. D. Are student ratings of teacher effectiveness valid? (IDEA Paper No. 2). Manhattan, KS: Center for Faculty Evaluation and Development, 1979.

Bausell, R. B., Schwartz, S., & Purohit, A. An examination of the conditions under which various student rating parameters replicate across time. Journal of Educational Measurement, 1975, 12, 273-280.

Braskamp, L. A., Caulley, D., & Costin, F. Student ratings and instructor self-ratings and their relationship to student achievement. American Educational Research Journal, 1979, 16, 295-306.

Braskamp, L., Ory, J. C., & Pieper, D. M. Student written comments: Dimensions of instructional quality. Journal of Educational Psychology, 1981, 73, 65-70.

Brown, D. L. Faculty ratings and student grades: A university-wide multiple regression analysis. Journal of Educational Psychology, 1976, 68, 573-578.

Cohen, P. A. Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. Research in Higher Education, 1980, 13, 321-341.

Feldman, K. A. Grades and college students' evaluations of their courses and teachers. Research in Higher Education, 1976, 4, 69-111.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. The generalizability of student ratings of instruction: Estimation of the teacher and course components. Journal of Educational Measurement, 1978, 15, 1-13.

Hogan, T. P. Similarity of student ratings across instructors, courses, and time. Research in Higher Education, 1973, 1, 149-154.

Howard, G. S., & Maxwell, S. E. Correlation between student satisfaction and grades: A case of mistaken causation? Journal of Educational Psychology, 1980, 72, 810-820.

Lamberth, J., & Kosteski, D. M. Student evaluations: An assessment of validity. Teaching of Psychology, 1981, 8, 8-11.

Marsh, H. W. The influence of student, course, and instructor characteristics in the evaluations of university teaching. American Educational Research Journal, 1980, 17, 219-237.

Marsh, H. W., Overall, J. U., & Kesler, S. P. Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluation and evaluations by their students. Journal of Educational Psychology, 1979, 71, 149-160.

Marsh, H. W., & Overall, J. U. The relative influence of course level, course type, and instructor on students' evaluations of college teaching. American Educational Research Journal, 1981, 18, 103-112.

Ory, J. C., Braskamp, L. A., & Pieper, D. M. Congruency of student evaluative information collected by three methods. Journal of Educational Psychology, 1980, 72, 181-185.

Overall, J. U., & Marsh, H. W. Midterm feedback from students: Its relationship to instructional improvement and student's cognitive and affective outcomes. Journal of Educational Psychology, 1979, 71, 856-865.

Overall, J. U., & Marsh, H. W. Students' evaluations of instructions: A longitudinal study of their stability. Journal of Educational Psychology, 1990, 72, 321-325.

Stumpf, S. A., Freedman, R. D., & Aguanno, J. C. A path analysis of factors often found to be related to student ratings of teaching effectiveness. Research in Higher Education, 1979, 11, 111-123.

Wise, S. L., Hengstler, D. D., & Braskamp, L. A. Alumni ratings as an indicator of departmental quality. Journal of Educational Psychology, 1981, 73, 71-77.

IDEA Papers may be ordered from the Center. Copies are 20¢ each, and the minimum order is 25 copies. Orders for less than $25.00 must be prepaid.

Center for Faculty Evaluation and Development, Kansas State University, 1623 Anderson Avenue, Manhattan, KS 66502-4098.